Q1) Identify the Data type for the Following:

|  |  |
| --- | --- |
| Activity | Data Type |
| Number of beatings from Wife | Discrete Data |
| Results of rolling a dice | Discrete Data |
| Weight of a person | Continues Data |
| Weight of Gold | Continues Data |
| Distance between two places | Continues Data |
| Length of a leaf | Continues Data |
| Dog's weight | Continues Data |
| Blue Color | Discrete Data |
| Number of kids | Discrete Data |
| Number of tickets in Indian railways | Discrete Data |
| Number of times married | Discrete Data |
| Gender (Male or Female) | Discrete Data |

Q2) Identify the Data types, which were among the following

Nominal, Ordinal, Interval, Ratio.

|  |  |
| --- | --- |
| Data | Data Type |
| Gender | Discrete data- Nominal |
| High School Class Ranking | Discrete data- Nominal |
| Celsius Temperature | Continuous- Interval |
| Weight | Continuous- Ratio |
| Hair Color | Discrete data- Ratio |
| Socioeconomic Status | Continuous- Interval |
| Fahrenheit Temperature | Continuous – Ratio |
| Height | Continuous- Ratio |
| Type of living accommodation | Discrete- Ordinal |
| Level of Agreement | Discrete- Interval |
| IQ(Intelligence Scale) | Discrete- Interval |
| Sales Figures | Discrete- Interval |
| Blood Group | Discrete- Ratio |
| Time Of Day | Continuous – Interval |
| Time on a Clock with Hands | Continuous- Interval |
| Number of Children | Discrete- Interval |
| Religious Preference | Discrete- Ratio |
| Barometer Pressure | Continuous – Interval |
| SAT Scores | Continuous- Ratio |
| Years of Education | Discrete- Nominal |

Q3) Three Coins are tossed, find the probability that two heads and one tail are obtained?

**Ans:** When 3 coins are tossed, their will be 8 possible outcomes.

Let ‘S’ be the Sample Space

S = {HHH, HHT, HTH, THH, TTT, TTH, THT, HTT}

Let ‘A’ be the event of getting two heads and one tail by tossing 3 coins

Therefore, Number of Favorable outcomes = 3

Total number of Possible outcomes = 8

P(A) = Number of Favorable outcomes / Total number of Possible outcomes

3/8 = 0.375

So that, the Probability of getting 2 Heads and 1 Tail is **0.375**

Q4) Two Dice are rolled, find the probability that sum is

1. Equal to 1
2. Less than or equal to 4
3. Sum is divisible by 2 and 3

**Ans:** When two dice are rolled, The possible outcomes is 6\*6 = 36

1. There will be no possible outcomes which is equal to 1,

P(A) = Number of Favorable outcomes / Total number of Possible outcomes

= 0/36 = 0

Therefore, Probability of getting outcome which is Equal to 1 is **0.**

1. There will be 6 possible outcomes of getting less than or equal to 4

i.e., S = {(1,1) (1,2) (1,3) (2,1) (2,2) (3,1)}

P(A) = 6/36

= 1/6

= 0.16

Therefore, Probability of getting outcome which is Less than or equal to 4 is **1/6 or 0.16**

1. There will be 6 possible outcomes of getting the Sum which is divisible by 2 and 3

i.e., S = {(1,5) (2,4) (3,3) (4,2) (5,1) (6,6)}

P(A) = 6/36

= 1/6

= 0.16

Therefore, Probability of getting the Sum which is divisible by 2 and 3 is **1/6 or 0.16**

Q5) A bag contains 2 red, 3 green and 2 blue balls. Two balls are drawn at random. What is the probability that none of the balls drawn is blue?

**Ans:** Total number of balls in a bag is 2+3+2 = 7

i.e., S = 7\*6/2 = 21 (Total number of Possible outcomes)

Number of drawing 2 balls out of (2+3) = 5

i.e., Number of Favorable outcomes = 10

P(A) = 10/21

Therefore, Probability of getting none of the balls drawn is blue is **10/21**

Q6) Calculate the Expected number of candies for a randomly selected child

Below are the probabilities of count of candies for children (ignoring the nature of the child-Generalized view)

|  |  |  |
| --- | --- | --- |
| CHILD | Candies count | Probability |
| A | 1 | 0.015 |
| B | 4 | 0.20 |
| C | 3 | 0.65 |
| D | 5 | 0.005 |
| E | 6 | 0.01 |
| F | 2 | 0.120 |

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

**Ans:** We know that,

Child A – probability of having 1 candy = 0.015.

Child B – probability of having 4 candies = 0.20

Expected number of candies for randomly selected child =

{(1\*0.015) + (4\*0.20) + (3\*0.65) + (5\*0.005) + (6\*0.01) + (2\*0.120)} = 3.09

Therefore, Expected number of candies for a randomly selected child

= **3.09**

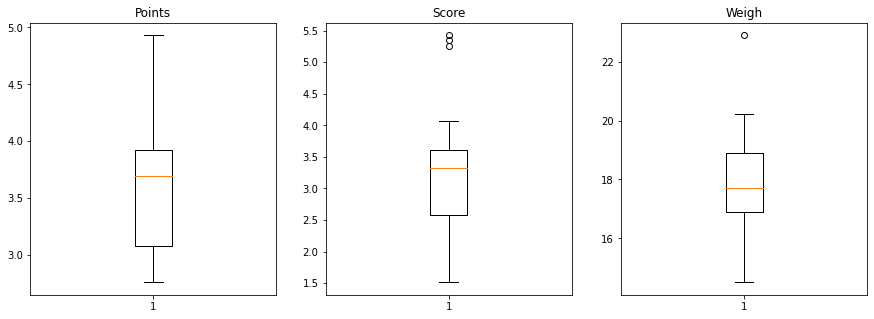
Q7) Calculate Mean, Median, Mode, Variance, Standard Deviation, Range & comment about the values / draw inferences, for the given dataset

* For Points,Score,Weigh>

Find Mean, Median, Mode, Variance, Standard Deviation, and Range and also Comment about the values/ Draw some inferences.

**Ans:**

|  |  |  |  |
| --- | --- | --- | --- |
|  | **Points** | **Score** | **Weigh** |
| Mean | 3.5965 | 3.2172 | 17.8487 |
| Median | 3.6950 | 3.325 | 17.71 |
| Mode | 3.92 | 3.44 | 18.90 |
| Variance | 0.2858 | 0.9573 | 3.1931 |
| Standard Deviation | 0.5346 | 0.9784 | 1.7869 |

****

**#CODES**

import numpy as np

import pandas as pd

import matplotlib.pyplot as plt

import seaborn as sns

%matplotlib inline

cars=pd.read\_csv("/content/Q7.csv")

**# Mean,median, variance, Standard Deviation for data set**

cars.Points.mean()

cars.Points.median()

cars.Points.mode()

cars.Points.var()cars.Points.std()

cars.Score.mean()

cars.Score.median()

cars.Score.mode()

cars.Score.var()

cars.Score.std()

cars.Weigh.mean()

cars.Weigh.median()

cars.Weigh.mode()

cars.Weigh.var()

cars.Weigh.std()

## Range ##

cars.describe()

Points\_Range=cars.Points.max()-cars.Points.min()

Points\_Range

Score\_Range=cars.Score.max()-cars.Score.min()

Score\_Range

Weigh\_Range=cars.Weigh.max()-cars.Weigh.min()

Weigh\_Range

f,ax=plt.subplots(figsize=(15,5))

plt.subplot(1,3,1)

plt.boxplot(cars.Points)

plt.title('Points')

plt.subplot(1,3,2)

plt.boxplot(cars.Score)

plt.title('Score')

plt.subplot(1,3,3)

plt.boxplot(cars.Weigh)

plt.title('Weigh')

plt.show()

Q8) Calculate Expected Value for the problem below

1. The weights (X) of patients at a clinic (in pounds), are

108, 110, 123, 134, 135, 145, 167, 187, 199

Assume one of the patients is chosen at random. What is the Expected Value of the Weight of that patient?

**Ans:** Number of patients = 9

Probability of selecting each patient = 1/9

Therefore,

P(X) = {(1/9\*108)+( 1/9\*110)+( 1/9\*123)+( 1/9\*134)+( 1/9\*135)+(

1/9\*145)+( 1/9\*167)+( 1/9\*187)+( 1/9\*199)}

= (1/9) \* (108+110+123+134+135+145+167+187+199)

= (1/9) \* (1308)

= 145.33

Hence, The Expected Value of the Weight of that patient is **145.33**

**Q9) Calculate Skewness, Kurtosis & draw inferences on the following data**

**Cars speed and distance**

**Ans:** The Skewness and Kurtosis of Speed and Dist is

|  |  |  |
| --- | --- | --- |
|  | **SPEED** | **DIST** |
| Skewness | -0.1139 | 0.7824 |
| Kurtosis | -0.5771 | 0.2480 |

The Skewness and Kurtosis of SP and WT

|  |  |  |
| --- | --- | --- |
|  | **SP** | **WT** |
| Skewness | 1.5814 | -0.6033 |
| Kurtosis | 2.7235 | 0.8194 |

**Q10) Draw inferences about the following boxplot & histogram**



**Ans:** The most of the data points are concerted in the range 50-100 with frequency 200.

And least range of weight is 400 somewhere around 0-10.

So the expected value the above distribution is 75.

Skewness- we can notice a long tail towards right so it is heavily right skewed.



**Ans:** Median is less than mean right skewed and we have outlier on the upper side of box plot and

there is less data points between Q1 and bottom point.

**Q11)** Suppose we want to estimate the average weight of an adult male in Mexico. We draw a random sample of 2,000 men from a population of 3,000,000 men and weigh them. We find that the average person in our sample weighs 200 pounds, and the standard deviation of the sample is 30 pounds. Calculate 94%,98%,96% confidence interval?

**Ans:** Confidence interval is 94%= (198.738325292158, 201.261674707842)

Confidence interval is 98%= (198.43943840429978, 201.56056159570022)

Confidence interval for 96%= (198.62230334813333, 201.37769665186667)

**#CODES**

**import** numpy **as** np

**import** pandas **as** pd

**from** scipy **import** stats

**from** scipy.stats **import** norm

stats**.**norm**.**interval(0.94,200,30**/**(2000**\*\***0.5))

stats**.**norm**.**interval(0.98,200,30**/**(2000**\*\***0.5))

stats**.**norm**.**interval(0.96,200,30**/**(2000**\*\***0.5))

**Q12)** Below are the scores obtained by a student in tests

**34,36,36,38,38,39,39,40,40,41,41,41,41,42,42,45,49,56**

1. Find mean, median, variance, standard deviation.
2. What can we say about the student marks?

**Ans:** Mean= 41

Median= 40.5

Variance= 25.5294

Standard deviation= 5.0526

Q13) What is the nature of skewness when mean, median of data are equal?

**Ans:** Symmetrical

Q14) What is the nature of skewness when mean > median ?

**Ans:** Right skewed

Q15) What is the nature of skewness when median > mean?

**Ans:** Left Skewed

Q16) What does positive kurtosis value indicates for a data ?

**Ans:** The data is normally distributed and kurtosis value is 0.

Q17) What does negative kurtosis value indicates for a data?

**Ans:** The distribution of the data has lighter tails and a flatter peaks than the normal distribution.

Q18) Answer the below questions using the below boxplot visualization.



What can we say about the distribution of the data?

**Ans:** Let’s assume above box plot is about a Hotel timings. 60% of the customers walking is above 10’o clock and Below 3’o clock. And the customers between 3’o clock to 6’o clock is approximately 40%.

What is nature of skewness of the data?

**Ans:** Left skewed, Median is greater than Mean.

What will be the IQR of the data (approximately)?   
**Ans:** Approximately= -8

Q19) Comment on the below Boxplot visualizations?



Draw an Inference from the distribution of data for Boxplot 1 with respect Boxplot 2.

**Ans:** By observing both the plots whisker’s level is high in boxplot 2, mean and median are equal hence distribution is symmetrical.

Q 20) Calculate probability from the given dataset for the below cases

Data \_set: Cars.csv

Calculate the probability of MPG of Cars for the below cases

MPG <- Cars$MPG

* 1. P(MPG>38)

**Ans:** 0.3475

* 1. P(MPG<40)

**Ans:** 0.7293

c. P (20<MPG<50)

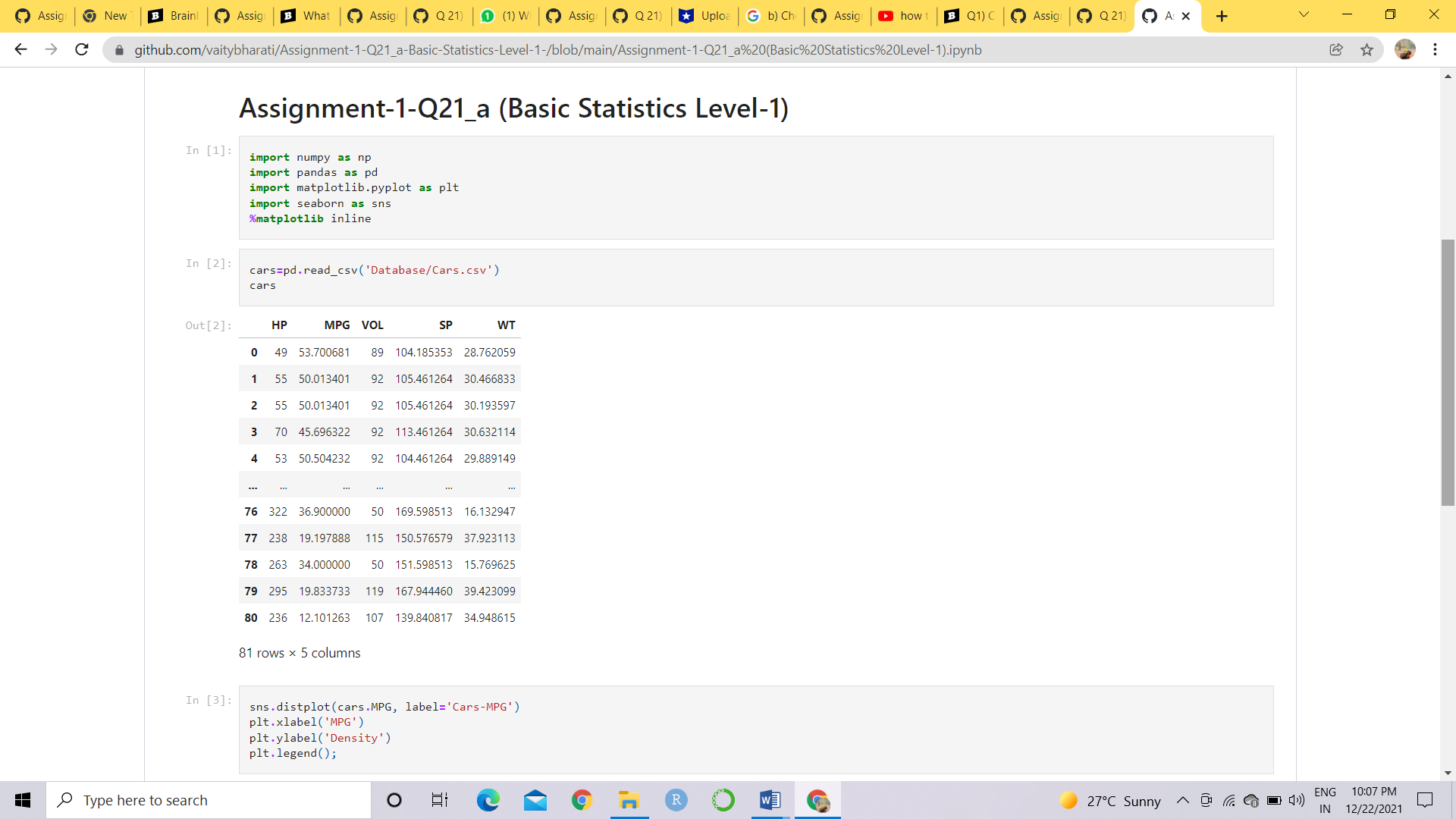
**Ans:** 1.2430

Q 21) Check whether the data follows normal distribution

1. Check whether the MPG of Cars follows Normal Distribution

Dataset: Cars.csv

**Ans:**



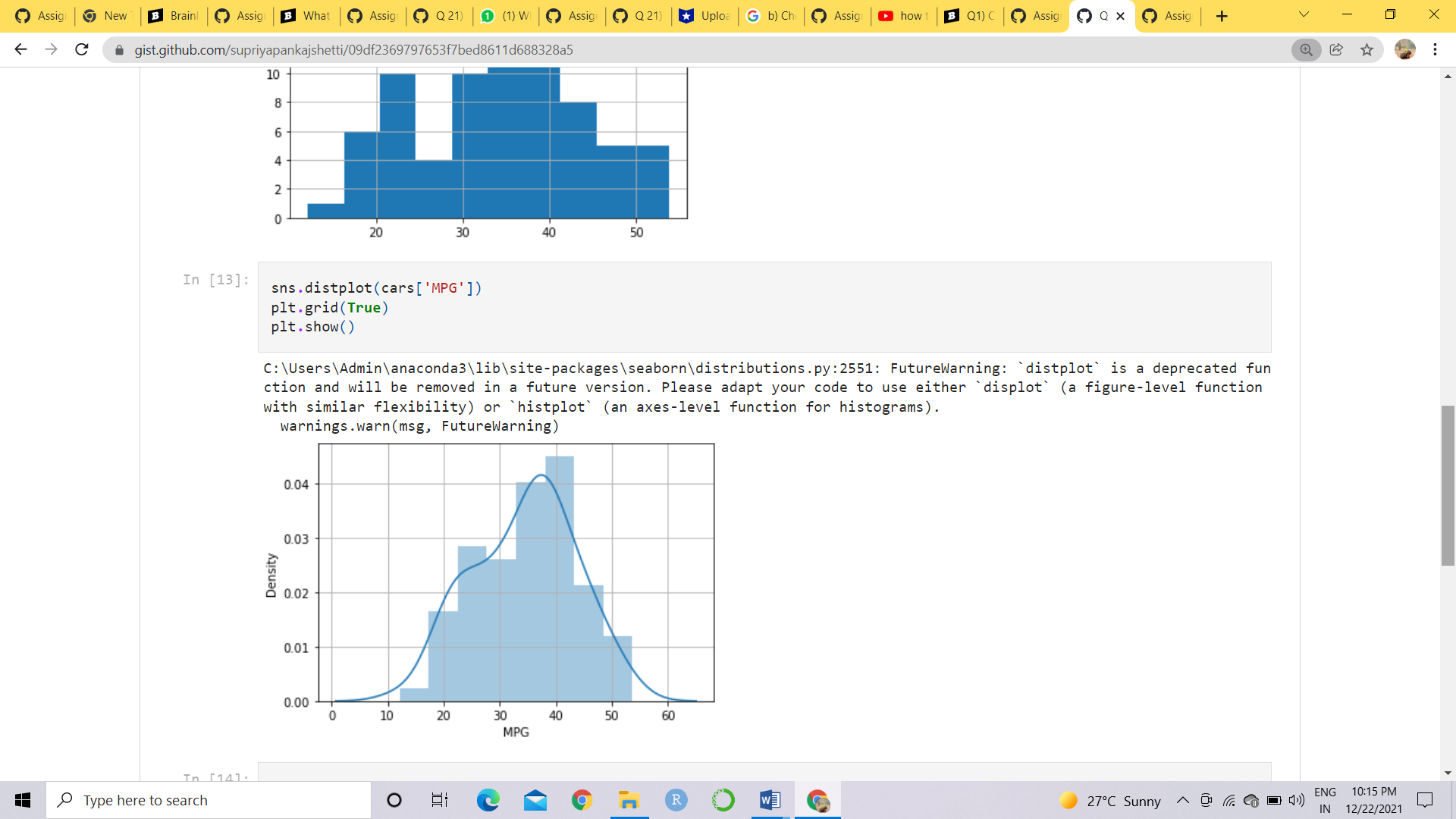
Mean = 34.4220

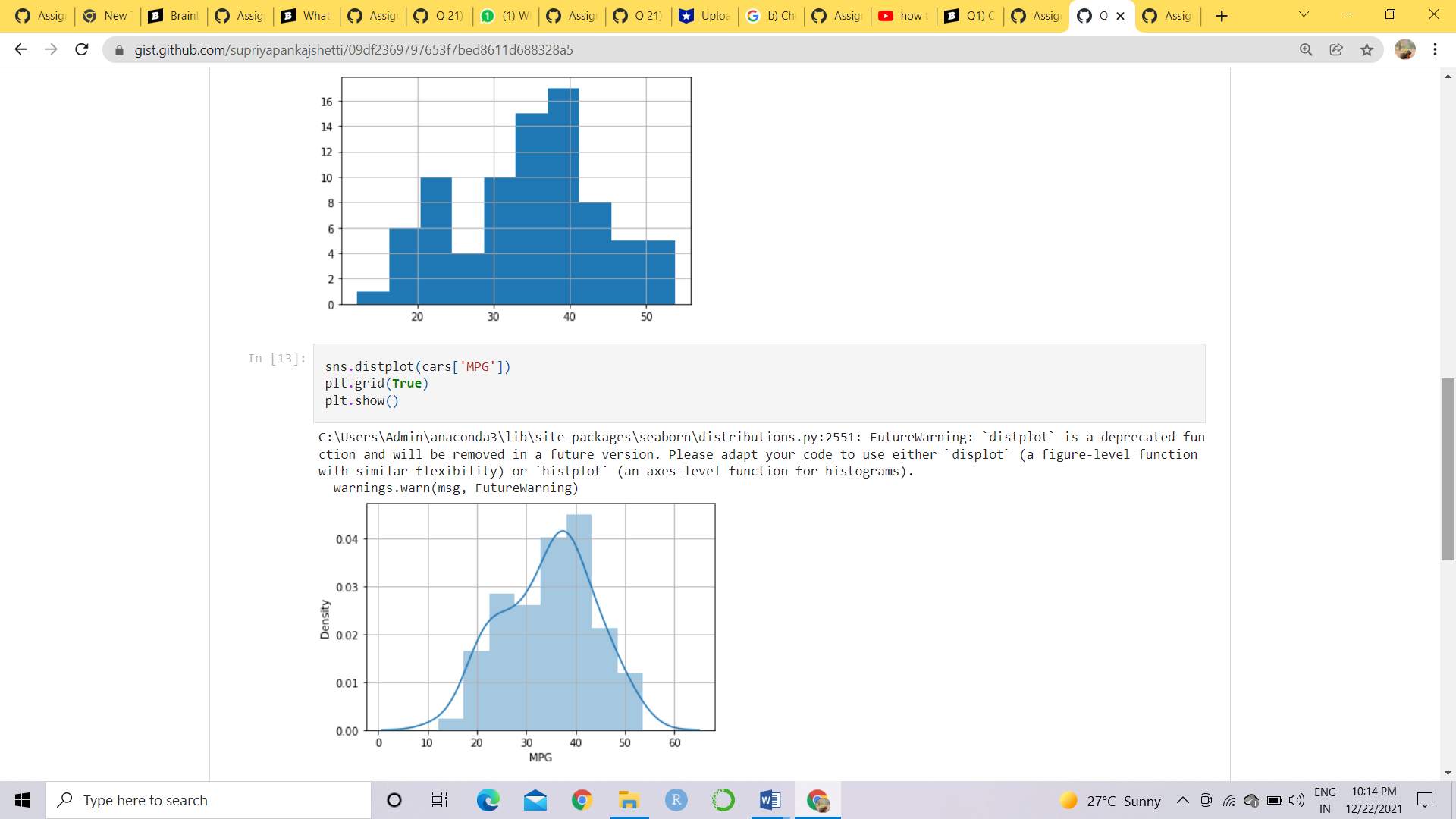
Median = 35.1527

Mode = 29.6299

Skewness = -0.1779

Kurtosis = -0.6116





From above plot and values we can say that data is **symmetrical** and **normally distributed**.

**#Codes**

**import** pandas **as** pd

**from** matplotlib **import** pyplot **as** plt

**import** seaborn **as** sns

cars**=**pd**.**read\_csv('/cars.csv')

cars

cars['MPG']**.**mean()

cars['MPG']**.**median()

cars['MPG']**.**mode()

cars['MPG']**.**hist()

sns**.**distplot(cars['MPG'])

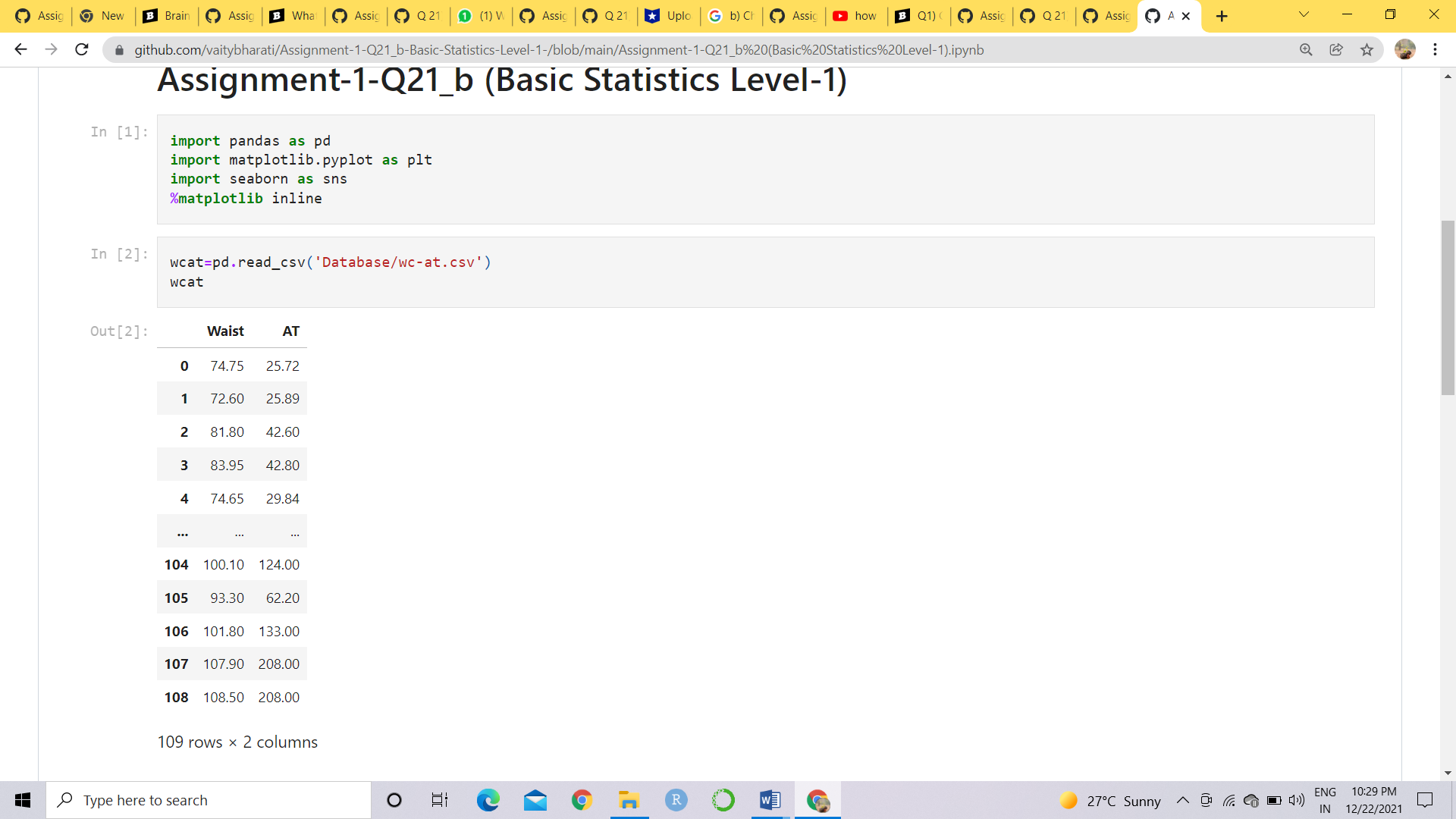
plt**.**grid(**True**)

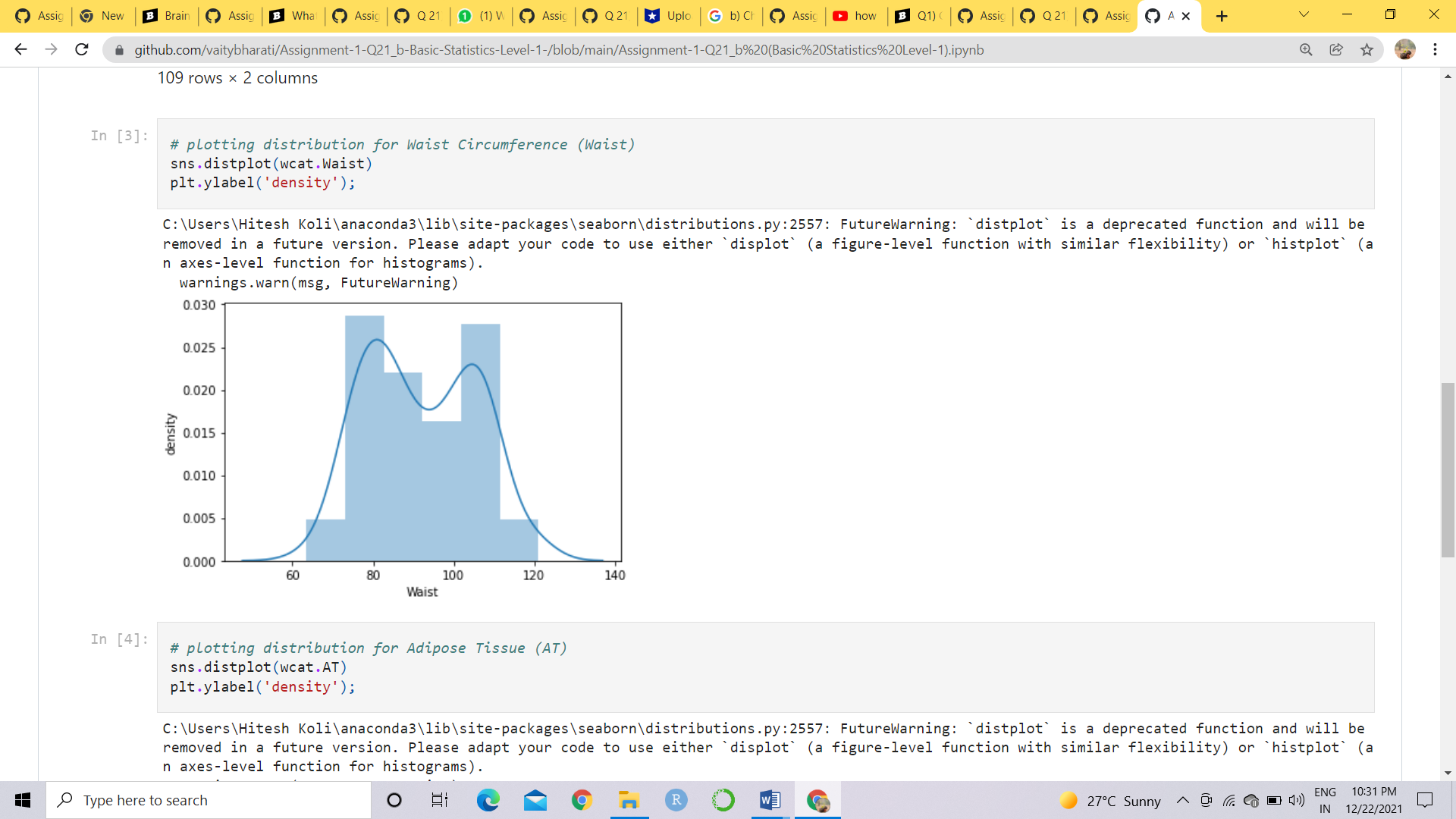
plt**.**show()

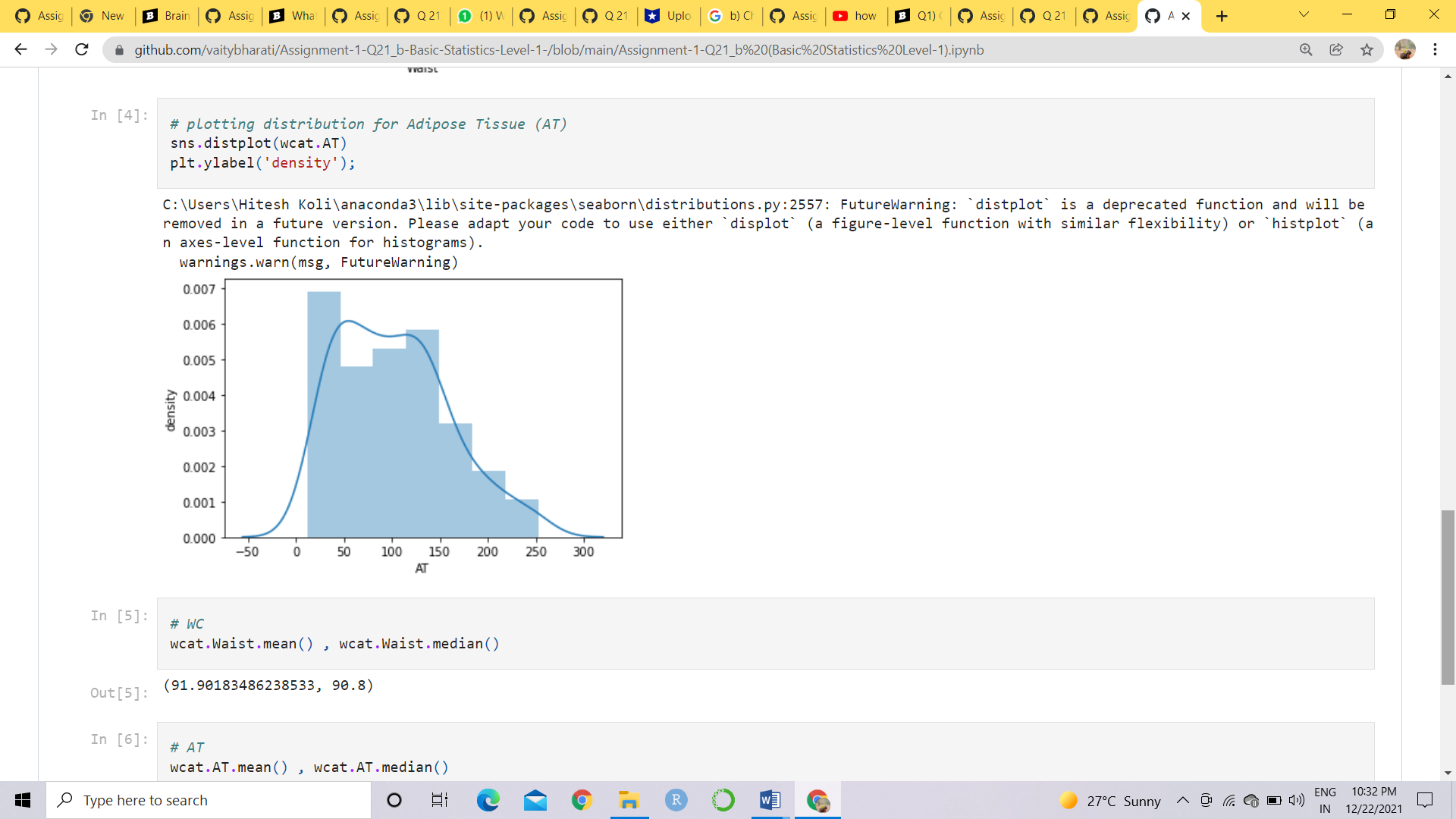
cars['MPG']**.**skew()

cars['MPG']**.**kurt()

1. Check Whether the Adipose Tissue (AT) and Waist Circumference(Waist) from wc-at data set follows Normal Distribution







WC MEAN = 91.90183486238533

WC MEDIAN = 90.8

AT MEAN =101.89403669724771

AT MEDIAN = 96.54

From above value and plots we can say that the data normally distributed

Q 22) Calculate the Z scores of 90% confidence interval,94% confidence interval, 60% confidence

interval

**Ans:** Z score for 90% confidence interval is 1.65

Z score for 94% confidence interval is 1.55

Z score for 60% confidence interval is 0.85

Q 23) Calculate the t scores of 95% confidence interval, 96% confidence interval, 99% confidence

interval for sample size of 25

**Ans:** The t scores for 95% confidence interval is 1.96

The t scores for 96% confidence interval is 2.5

The t scores for 99% confidence interval is 2.47

Q 24**)** A Government company claims that an average light bulb lasts 270 days. A researcher randomly selects 18 bulbs for testing. The sampled bulbs last an average of 260 days, with a standard deviation of 90 days. If the CEO's claim were true, what is the probability that 18 randomly selected bulbs would have an average life of no more than 260 days

**Ans:**

**Given,**

x = mean of the sample of bulbs = 260

μ = population mean = 270

s = standard deviation of the sample = 90

n = number of items in the sample = 18

WKT: t = x- μ/8/√n

t = 260 – 270/90/√18

t = -10/90/3√2

t = -10/30/√2

t = -1 \* √2 / 3

t = -0.471

Probability calculations, the number of degree of freedom is n - 1, so here you need the t-distribution with 17 degrees of freedom.

The probability that t < - 0.471 with 17 degrees of freedom assuming the population mean is true, the t-value is less than the t-value obtained With 17 degrees of freedom and a t score of - 0.471, the probability of the bulbs lasting less than 260 days on average of 0.3218 assuming the mean life of the bulbs is 300 days.

#CODES

**from** scipy **import** stats

**from** scipy.stats **import** norm

*Null Hypothesis is: Ho = Avg life of Bulb >= 260 days*

*Alternate Hypothesis is: Ha = Avg life of Bulb < 260 days*

*t-scores at x=260; t=(s\_mean-P\_mean)/(s\_SD/sqrt(n))*

t**=**(260**-**270)**/**(90**/**18**\*\***0.5)t

-0.4714045207910317

*Find P(X>=260) for null hypothesis*

*p\_value=1-stats.t.cdf(abs(t\_scores),df=n-1)... Using cdf function*

p\_value**=**1**-**stats**.**t**.**cdf(abs(**-**0.4714),df**=**17)

p\_value

0.32167411684460556

*# OR p\_value=stats.t.sf(abs(t\_score),df=n-1)... Using sf function*

p\_value**=**stats**.**t**.**sf(abs(**-**0.4714),df**=**17)

p\_value

0.32167411684460556